Task description:

Task: Please do some analysis on the file https://dl.dropboxusercontent.com/u/68829208/sequenceFile.zip and write a report on your findings. The report should contain at least evidences to answer the questions and could also contain anything that you find interesting on the file.

Notes:

1. This file contains NGS reads for some human DNA regions. In the zipped file there are the raw fastq reads file and also the alignment file in BAM format. The fastq file format is described at http://en.wikipedia.org/wiki/FASTQ_format. The bam file format is described at http://samtools.sourceforge.net/SAM1.pdf. raw file is aligned to hg19 human genome.
2. To access the content of the alignment BAM file, you could refer to samtools. (http://samtools.sourceforge.net)
3. The sequenced file is pair-ended from Miseq sequencing platform.
4. This file is the result of an amplicon sequencing design.
5. The primers used in this file are:

Forward Primer	Reverse Primer
TTGCCAGTTAACGTCTTCCTTCTCTCTCTG	GAGAAAAGGTGGGCCTGAGGTTCAGAGCCA
CCCTTGTCTCTGTGTTCTTGTCCCCCCCA	CCCCACCAGACCATGAGAGGCCCTGCGGCC
TGATCTGTCCCTCACAGCAGGGTCTTCTCT	TGACCTAAAGCCACCTCCTTA
CACACTGACGTGCCTCTCCCTCCCTCCA	CCGTATCTCCCTTCCCTGATTA

6. The adaptor sequences used in this file are:

Adaptor 1	AAGACTCGGCAGCATCTCCA
Adaptor 2	GCGATCGTCACTGTTCTCCA

Questions:

1. What is the constitution of each read, for example (adaptor + primer + amplified region) ?
2. Which gene region did those reads mapped to ?
3. What is the percentage of reads that were mapped to the human genome? Can you give some comments on the unmapped reads?
4. What is the coverage of each amplicons? Could you find any explanations or clues why the coverage varies among amplicons, especially for those with big differences?
5. What variants/mutations can you identify from this data? How can you evaluate which variants is more real than others?

6. The above questions are intended to provide you with some directions you could look at on the data and you are encouraged to explore what ever aspects you find interesting.

Please finish the task independently and hand in the report and program code within 1 week. Please be aware that some of the questions are open and you are also encouraged to explore one of the questions more deeply. Creative hypothesis and insights in understanding of the data are certainly a plus. The report should be in PDF format and figures rather than descriptive words are preferred. Please also make clear description on your figures.